Some Basic Concepts of Statistical Inference (Sec 5.1)

Suppose we have a rv X that has a pdf/pmf denoted by $f(x;\theta)$ or $p(x;\theta)$, where θ is called the parameter. In previous lectures, we focus on probability problems where the value of θ is given. From now on, we focus on statistical problems where θ is unknown and we try to get some information about θ from a random sample $(X_1,\ldots,X_n)$ from this distribution.

- Parameter θ
- Random sample: $(X_1,\ldots,X_n)$ iid $\sim f(\cdot;\theta)$
- Observed sample: $(x_1,\ldots,x_n)$ one realization of $(X_1,\ldots,X_n)$
- Statistic $T=T(X_1,\ldots,X_n)$: a function of the sample, which is also random.
- Estimator $\hat\theta=\hat\theta(X_1,\ldots,X_n)$ of θ is a function of the sample, i.e., a statistic. Given an observed sample $(X_1=x_1,\ldots,X_n=x_n)$, the value of $\hat\theta(x_1,\ldots,x_n)$ is called an Estimate of θ. So, an estimator is a random variable, while an estimate is a real number (i.e., one realization of the Estimator).

Overview of Estimation

- How to derive an estimator?
  - Method of Moments: suppose $\mathbb{E}(X)=h(\theta)$.

    $$\text{Set the sample mean} \bar X = h(\tilde\theta), \quad \text{then solve for } \tilde\theta.$$

  - Maximum Likelihood Estimator (see below).

  Notation: I'll use $\hat\theta$ as a generic notation for an estimator of parameter θ. In problems where we need derive estimators based on different approaches, I use $\tilde\theta$ for one estimator of θ and $\hat\theta$ another estimator of θ, for example, $\tilde\theta$ for method of moments estimator and $\hat\theta$ for MLE.

- How to evaluate the performance of an estimator?
  - Note that $\hat\theta$ is a random variable, usually a continuous random variable, so the chance that $\hat\theta=\theta$ is zero. But we can say whether *on average* $\hat\theta$ is equal to θ, which leads to the definition of

    $$\text{Bias}(\hat\theta)=\mathbb{E}(\hat\theta)-\theta.$$

  - Another metric is the *averaged* square distance from $\hat\theta$ to the target θ, which leads to the definition of Mean Squared Error

    $$\text{MSE}(\hat\theta)=\mathbb{E}\big[(\hat\theta-\theta)^2\big].$$

– If we have derived multiple estimators for θ, we can compare their MSE's by their relative efficiency.

Maximum Likelihood Estimator (MLE, Sec 6.1)

MLE: the estimator or estimators[1] that maximize the Likelihood function

$$L(\theta; \mathbf{x}) = f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

How to derive MLE?

Step 1: Compute $\log f(x; \theta)$;

Step 2: Plug x_i into the expression derived at Step 1 and sum them over i, which gives us the log likelihood function:

$$\ell(\theta) = \log \Big[\prod_{i=1}^{n} f(x_i; \theta) \Big] = \sum_{i=1}^{n} \log f(x_i; \theta).$$

Step 3: Find the maximum of $\ell(\theta)$: $\hat{\theta} = \arg\max_\theta \ell(\theta)$.

Tips/Tools for Step 3.

- Take derivative of $\ell(\theta)$ with respect to θ and solve $\ell'(\theta) = 0$ for θ.

  Be careful when the parameter θ is in a bounded region, say $\theta \geq 0$ or $\theta \in [0, 1]$. Make sure the solution $\hat{\theta}$ is in the range of θ. If it's outside the range, usually you need to check whether one of the boundary points is the maximum.

- If X is bounded and the bounds depend on θ, remember to add the indicator function in $f(x; \theta)$. For example, if $X \sim \mathsf{Unif}(0, \theta]$, then $f(x; \theta) = \frac{1}{\theta} \mathbf{1}(0 \leq x \leq \theta)$.

- Some special optimization results (proofs are in the Appendix):
  - $\min_a \sum_{i=1}^{n} |x_i - a|$ is achieved at $a = \text{median}(x_i)$.
  - $\max_{p_1, \ldots, p_m} \sum_{j=1}^{m} n_j \log p_j$ where $0 \leq p_j \leq 1$ and $\sum_j p_j = 1$ and $n_j \geq 0$, is achieved by setting $p_j = n_j / n$ where $n = n_1 + \cdots + n_m$.

- The objective function could have multiple solutions or no solution (i.e., MLE doesn't exist).

Thm 6.1.2 (*Invariance property of MLE*) Let $X_1, \ldots, X_n$ be a random sample with the pdf $f(x; \theta)$. Let $\eta = g(\theta)$ be a parameter of interest. Suppose $\hat{\theta}$ is the mle of θ. Then $g(\hat{\theta})$ is the mle of $g(\theta)$.

[1] MLE may not be unique.

Below we list the MLE the method of moments (MM) estimators for various distribution families. The derivation isn't difficult. In class, I'll go through some of them, but expect you to go through the remaining by yourself.

Dist	pmf/pdf	MLE	MM
$\mathsf{Bern}(\theta)$	$\theta^x(1-\theta)^{1-x}$	$\bar{X}$	$\bar{X}$
$\mathsf{Bin}(n,\theta)$	$\binom{n}{y}\theta^y(1-\theta)^{n-y}$	Y/n	Y/n
$\mathsf{Geo}(p)$	$p(1-p)^{(x-1)}$	$1/\bar{X}$	$1/\bar{X}$
$\mathsf{Po}(\lambda)$	$\frac{\lambda^x}{x!}e^{-\lambda x}$	$\bar{X}$	$\bar{X}$
$\mathsf{Ex}(\lambda)$	$\lambda x^{-\lambda x}$	$1/\bar{X}$	$1/\bar{X}$
$\mathsf{Ex}(1/\theta)$	$\frac{1}{\theta}x^{-x/\theta}$	$\bar{X}$	$\bar{X}$
$\mathsf{N}(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/(2\sigma^2)}$	$\bar{X}$, $\frac{1}{n}\sum_i(X_i-\bar{X})^2$	$\bar{X}$, $\frac{1}{n-1}\sum_i(X_i-\bar{X})^2$
$\mathsf{Beta}(\alpha,1)$	$\frac{\Gamma(\alpha+1)}{\Gamma(\alpha)}x^{\alpha-1}$	$n/(\sum_i \log\frac{1}{X_i})$	$\frac{\bar{X}}{1-\bar{X}}$
$\mathsf{Beta}(1/\theta,1)$	$\frac{\Gamma(\frac{1}{\theta}+1)}{\Gamma(\frac{1}{\theta})}x^{1/\theta-1}$	$\frac{\sum_i \log\frac{1}{X_i}}{n}$	$\frac{1}{\bar{X}}-1$
$\mathsf{Ga}(\alpha,\beta)$, α known	$\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$	$\alpha/\bar{X}$	$\alpha/\bar{X}$
$\mathsf{Ga}(\alpha,1/\theta)$, α known	$\frac{1}{\Gamma(\alpha)\theta^\alpha}x^{\alpha-1}e^{-x/\theta}$	$\bar{X}/\alpha$	$\bar{X}/\alpha$

Table 1: List of MLE and MM estimators for various parametric families.

- For the binomial distribution, we consider the case where we have only one observation $Y \sim \mathsf{Bin}(n,\theta)$. If we have $Y_1,\ldots,Y_m$ iid $\sim \mathsf{Bin}(n,\theta)$, then the MLE and MM estimator will be $\bar{Y}/n = (Y_1+\cdots+Y_m)/(mn)$.
- The parameterization of $\mathsf{Geo}(p)$ is different from the one in Appendix D. Here $X \sim \mathsf{Geo}(p)$ denotes the number of Bernoulli trials you have conducted before seeing the first Head, including the last trial in which you observe a Head. That is, $X = 1, 2, \ldots$ and you observe one Head and $(X-1)$ Tails.
- $\mathsf{Beta}(1/\theta,1)$ is discussed on p3 of `Week7_Estimation1ans.pdf`.

More Examples

- (6.1.3): Let $X_1, \ldots, X_n$ be a random sample from the Laplace distribution with pdf

$$f(x;\theta) = \frac{1}{2}e^{-|x-\theta|}.$$

Show that the mle of θ is given by $\hat{\theta} = \text{median}(X_1, \ldots, X_n)$.

Since the mean of this distribution is θ, the method of moments estimator should be $\bar{X}$.

- (6.4.5): Suppose we have a bag of marbles in three different colors, red, blue, and green. To estimate the proportion of the three colors (p_1, p_2, p_3), we conducted the following experiment: we randomly draw a marble from the bag, record its color,

$$X_i = \begin{cases} 1, & \text{if red} \\ 2, & \text{if blue} \\ 3, & \text{if green} \end{cases}$$

and put it back to the bag; repeat this process n times. X_i's are iid samples from a multinomial distribution: $X_i = 1, 2, 3$ with probabilities p_1, p_2, p_3 respectively.

The joint likelihood of $X_1, \ldots, X_n$ is equal to

$$f(X_1, \ldots, X_n \mid p_1, p_2, p_3) = p_1^{n_1} \times p_2^{n_2} \times p_3^{n_3},$$

which only depends on n_j = number of j's we have in the n marbles. To find the MLE of $\mathbf{p} = (p_1, p_2, p_3)^t$, we can maximize the log-likelihood function

$$\ell(\mathbf{p}) = n_1 \log p_1 + n_2 \log p_2 + n_3 \log p_3 \tag{1}$$

subject to the constraints that

$$0 \le p_1, p_2, p_3 \le 1 \quad \text{and} \quad p_1 + p_2 + p_3 = 1.$$

The MLE is given by

$$\hat{p}_1 = \frac{n_1}{n}, \quad \hat{p}_2 = \frac{n_2}{n}, \quad \hat{p}_3 = \frac{n_3}{n},$$

i.e., we use the sample frequencies to estimate the proportion. The MLE and MM estimator are the same.

If we have observed 2 red marbles, 3 blue marbles, and 5 green marbles, then the MLE for (p_1, p_2, p_3) is given by

$$\hat{p}_1 = .2, \quad \hat{p}_2 = .3, \hat{p}_3 = .5$$

- Let $X_1, \ldots, X_n$ be a random sample from $\mathsf{Ex}(\lambda)$ with pdf

$$f(x;\theta) = \lambda x^{-\lambda x}, \quad x > 0.$$

a) (6.1.2) Show that the mle of λ is given by $\hat{\lambda} = 1/\bar{X}$.

b) What's the mle of the probability $\mathbb{P}(X > 1)$?

c) If we parameterize the Exponential family by $\theta = 1/\lambda$, what's the mle of θ?

- Let $X_1, \ldots, X_n$ be a random sample from $\mathsf{Bern}(\theta)$ with pmf

$$p(x;\theta) = \theta^x (1-\theta)^x, \quad x = 0, 1.$$

a) (6.1.1) Show that the mle of θ is given by $\hat{\theta} = \bar{X}$.

b) Let $Y = X_1 + \cdots + X_n$. So $Y \sim \mathsf{Bin}(n, \theta)$. Derive the mle of θ given Y.

c) (6.1.6) Show that the MLE of θ is given by $\hat{\theta} = \min(\bar{X}, \frac{1}{3})$, if $0 \le \theta \le 1/3$.

- (Example 1 from `Week8_Estimation2ans.pdf`) Let λ and let $X_1, \ldots, X_n$ be a random sample from the distribution with pdf

$$f(x;\lambda) = 2\lambda^2 x^3 e^{-\lambda x^2}, \quad x > 0.$$

Define $Y = X^2$, which is a one-to-one transformation since $X > 0$. The pdf for Y is given by

$$\begin{aligned} f_Y(y) &= f_X(\sqrt{y}) \left| \frac{dy}{dx} \right|, \qquad \text{where } \frac{dx}{dy} = \frac{d\sqrt{y}}{dy} = \frac{1}{2\sqrt{y}} \\ &= 2\lambda^2 \sqrt{y}^{3/2} e^{-\lambda y} \frac{1}{2\sqrt{y}} \\ &= y e^{-\lambda y} \end{aligned}$$

i.e., $Y \sim \mathsf{Ga}(2, 1/\lambda)$. So the MLE and the MM estimator of λ are the same, given by

$$2/\bar{Y} = \frac{2}{\frac{1}{n}\sum_i X_i^2} = \frac{2n}{\sum_i X_i^2}.$$

- (6.1.5): Let $X_1, \ldots, X_n$ be a random sample from $\mathsf{Unif}(0, \theta]$ with pdf

$$f(x;\theta) = \frac{1}{\theta} I_{(0<x\le\theta)}.$$

a) Derive the MLE of θ.

The likelihood function is given by

$$f(x_1, \ldots, x_n) = \prod_{j=1} n \frac{1}{\theta} I_{(0<x_i\le\theta)} = \frac{1}{\theta^n} I_{(0< \text{ all } x_i\text{'s } \le\theta)}.$$

Graph this function: it's a monotone decreasing function of θ when $\theta \ge \max_i x_i$, but zero when $\theta < \max_i x_i's$. So $\hat{\theta} = \max_i X_i$.

b) Derive the MM of θ. $\mathbb{E}X = 2\theta$, so $\tilde{\theta} = \bar{X}/2$.

c) What if we change the distribution to be $\mathsf{Unif}(0, \theta)$, i.e.,

$$f(x;\theta) = \frac{1}{\theta} I_{(0<x<\theta)}.$$

The likelihood function becomes $\frac{1}{\theta^n} I_{(0< \text{ all } x_i\text{'s } <\theta)}$, which is a decreasing function when $\theta > \max_i x_i$, but zero when $\theta \le \max_i x_i$. So the MLE does not exist in this case.

Unbiased Estimators

An estimator is called unbiased if $\mathbb{E}(\hat{\theta}) = \theta$. The following results are useful when checking whether an estimator is biased/unbiased.

$$\begin{aligned} \mathbb{E}(\sum_i a_i X_i) &= a_i \mathbb{E}(X_i) \\ \mathrm{Var}(\sum_i a_i X_i) &= \sum_{i,j} a_i a_j \mathrm{Cov}(X_i, X_j) = \sum_i a_i^2 \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j). \end{aligned}$$

- If $\mathbb{E}(X) = \theta$, and your estimator of $g(\theta)$ is $g(\bar{X})$, then likely you need to use Jensen's inequality to show that $g(\bar{X})$ is biased. You need to check whether g or $-g$ is convex, and then call the Jensen's inequality.

$$\begin{aligned} g''(x) \ge 0 &: \ g \text{ is convex and } \mathbb{E}g(\bar{X}) \ge g(\theta); \\ g''(x) \le 0 &: \ -g \text{ is convex and } \mathbb{E}g(\bar{X}) \le g(\theta); \end{aligned}$$

- Let $(X_1, \ldots, X_n)$ be a random sample from a distribution with mean μ and variance σ^2. Then

$$\mathbb{E}\bar{X} = \mu, \quad \mathbb{E}(S^2) = \mathbb{E}\Big[\frac{1}{n-1}\sum_{i=1}^n \big(X_i - \bar{X}\big)^2\Big] = \sigma^2.$$

That is, sample mean and sample variance are unbiased. How to show $\mathbb{E}(S^2) = \sigma^2$?

$$\begin{aligned} \sum_{i=1}^n \big(X_i - \bar{X}\big)^2 &= \sum_{i=1}^n \big(X_i^2 - 2X_i \cdot \bar{X} + \bar{X}^2\big) \\ &= \Big(\sum_{i=1}^n X_i^2\Big) - 2\bar{X}\Big(\sum_{i=1}^n X_i\Big) + \Big(\sum_{i=1}^n \bar{X}^2\Big) \\ &= \Big(\sum_{i=1}^n X_i^2\Big) - 2\bar{X} \cdot n\bar{X} + n\bar{X}^2 \\ &= \sum_{i=1}^n X_i^2 - n\bar{X}^2. \end{aligned}$$

$$\begin{aligned}
\mathbb{E}\Big[\frac{1}{n-1}\sum_{i=1}^n (X_i-\bar{X})^2\Big] &= \frac{1}{n-1}\mathbb{E}\Big[\sum_{i=1}^n X_i^2 - n\bar{X}^2\Big] \\
&= \frac{1}{n-1}\Big[\sum_{i=1}^n \mathbb{E}(X_i^2) - n\mathbb{E}(\bar{X}^2)\Big] \\
&= \frac{1}{n-1}\Big[\sum_{i=1}^n (\mu^2+\sigma^2) - n(\mu^2+\frac{\sigma^2}{n})\Big] \\
&= \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2.
\end{aligned}$$

- Most estimators in Table 1 are unbiased, except
  - $\mathsf{Ex}(\lambda)$. The estimator $1/\bar{X}$ is biased (Jensen's inequality), and the estimator for $\theta = 1/\lambda$ are unbiased.
  - $\mathsf{No}(\mu, \sigma^2)$. The MLE of σ^2 is biased.
  - $\mathsf{Geo}(p)$. The estimator $1/\bar{X}$ is biased (Jensen's inequality).
  - $\mathsf{Beta}(\alpha, 1)$. Both the MLE and the MM estimator of α are biased (Jensen's inequality). The MLE of $\theta = 1/\alpha$ is unbiased, but the MM estimator of θ is biased (Jensen's inequality). As we'll say that this case is the same as the exponential distribution since $-\log X$ follows an Exponential distribution.
  - $\mathsf{Ga}(\alpha, \beta)$ with α known. The estimator $\alpha/\bar{X}$ for β is biased ((Jensen's inequality), but the estimator $\bar{X}/\alpha$ for $\theta = 1/\beta$ is unbiased.

More Examples.

- Let $(X_1, \ldots, X_n)$ be a random sample from $\mathsf{Unif}(0, \theta]$. The MLE is biased.

$$Y_n = \max_i X_i, \quad \mathbb{E}(Y_n) = \frac{n}{n+1}\theta.$$

  The MM estimator $\tilde{\theta} = 2\bar{X}$ is unbiased.

- Let $(X_1, \ldots, X_n)$ be a random sample from a distribution with pdf $f(x)$ whose mean μ exists and which is *symmetric* about μ. Show that

$$\mathbb{E}(\text{sample median}) = \mu.$$

  Without loss of generality, assume $\mu = 0$ (Why?). That is, the pdf $f(x)$ is symmetric about zero. Due to the symmetry, we can show that

- The distribution of a random sample $(X_1, \ldots, X_n)$ should be the same as the distribution of $(-X_1, \ldots, -X_n)$.
- So the distribution of the sample median of $(X_1, \ldots, X_n)$ should be the same as the distribution of the sample median of $(-X_1, \ldots, -X_n)$.

$$\begin{aligned} \mathbb{E}(\text{sample median of } X_{1:n}) &= \mathbb{E}(\text{sample median of } -X_1, \ldots, -X_n) \\ &= -\mathbb{E}(\text{sample median of } X_{1:n}). \end{aligned}$$

So $\mathbb{E}(\text{sample median of } X_{1:n}) = 0$.

$X_1, \ldots, X_n \sim$ Laplacian distribution with pdf $f(x;\theta) = \frac{1}{2}e^{-|x-\theta|}$. The MLE $\hat{\theta} = \text{Med}(X_1, \ldots, X_n)$ is unbiased.

Mean Squared Error (MSE)

- For an estimator $\hat{\theta}$ of θ, define the Mean Squared Error of $\hat{\theta}$ by

$$\text{MSE}(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta)^2 = \mathbb{E}[\hat{\theta} - \mathbb{E}(\hat{\theta})]^2 + \text{Var}(\hat{\theta}) = \text{Bias}^2 + \text{Var}$$

Specially, if $\hat{\theta}$ is unbiased, then $\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta})$.

- Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two unbiased estimator of $\hat{\theta}$. $\hat{\theta}_1$ is said to be more efficient than $\hat{\theta}_2$ if

$$\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2).$$

The relative efficiency of $\hat{\theta}_1$ with respect to (wrt) $\hat{\theta}_2$ is $\text{Var}(\hat{\theta}_2)/\text{Var}(\hat{\theta}_1)$.

Examples.

- $X_1, \ldots, X_n \sim \mathsf{Unif}(0, \theta]$. We have learned that the MLE $\hat{\theta}$ and the MM estimator $\tilde{\theta}$ are given by

$$\hat{\theta} = Y_n = \max_i X_i, \quad \tilde{\theta} = 2\bar{X}.$$

  - Which estimator is "better" (i.e. having the smallest MSE), $\hat{\theta}$ or $\tilde{\theta}$?

$$\text{MSE}(\hat{\theta}) = \frac{2\theta^2}{(n+1)(n+2)}, \quad \text{MSE}(\tilde{\theta}) = \frac{\theta^2}{3n}.$$

$$
\begin{aligned}
f_{Y_n}(y) &= n\frac{y^{n-1}}{\theta^n}, \quad 0 < y \le \theta. \\
\mathbb{E}Y_n &= \int_0^\theta y f_{Y_n}(y)dy = \int_0^\theta n\frac{y^n}{\theta^n} = \frac{n}{n+1}\theta \\
\mathbb{E}Y_n^2 &= \int_0^\theta y^2 f_{Y_n}(y)dy = \int_0^\theta n\frac{y^{n+1}}{\theta^n} = \frac{n}{n+2}\theta^2 \\
\mathrm{Var}(Y_n) &= \mathbb{E}Y_n^2 - (\mathbb{E}Y_n)^2 = \frac{n}{(n+1)^2(n+2)}
\end{aligned}
$$

$$
\mathrm{MSE}(\hat\theta) = \mathrm{Bias}^2 + \mathrm{Var}(Y_n) = (\mathbb{E}Y_n - \theta)^2 + \mathbb{E}Y_n^2 - (\mathbb{E}Y_n)^2.
$$

$$
\mathrm{MSE}(\tilde\theta) = 4\mathrm{Var}(\bar X) = \frac{4}{n}\mathrm{Var}(X_1) = \frac{4}{n}\times\frac{\theta^2}{12} = \frac{\theta^2}{3n}
$$

Note that although $\tilde\theta = 2\bar X$ is unbiased for θ, while $\hat\theta = \max_i X_i$ is biased. The MSE of $\hat\theta$ is much smaller than the MSE of $\tilde\theta$ for large n.

– What must c equal if $c\hat\theta$ is to be an unbiased estimator of θ? That is, construct an unbiased estimator based the MLE $\hat\theta$.

– Which estimator is more efficient, $\tilde\theta$ or $\frac{n+1}{n}\hat\theta$? What's the relative efficiency of $\frac{n+1}{n}\hat\theta$ wrt $\tilde\theta$?

$$
\begin{aligned}
\mathrm{MSE}\Big(\frac{n+1}{n}\hat\theta\Big) &= \mathrm{Var}\Big(\frac{n+1}{n}\hat\theta\Big) = \frac{(n+1)^2}{n^2}\mathrm{Var}(Y_n) \\
&= \frac{(n+1)^2}{n^2}\frac{n\theta^2}{(n+1)^2(n+2)} = \frac{\theta^2}{n(n+2)}.
\end{aligned}
$$

The relative efficiency of $\frac{n+1}{n}\hat\theta$ wrt $\tilde\theta$ is equal to

$$
\frac{\theta^2}{3n}\Big/\frac{\theta^2}{n(n+2)} = \frac{n+2}{3}.
$$

When the sample size n gets larger, the estimator $\frac{n+1}{n}\hat\theta$ (that is also unbiased) is much more efficient than the MM estimator $\tilde\theta$.

(*) (Sec 7.1) In general, it's difficult to compare two estimators based on their MSE, since MSE may depend on the unknown parameter θ: It is possible that $\mathrm{MSE}(\hat\theta_1) < \mathrm{MSE}(\hat\theta_2)$ (i.e., $\hat\theta_1$ is better) when $\theta > 1$, while $\hat\theta_2$ is better for $\theta < 1$. In courses on statistical decision theory, you'll learn how to compare estimators based on their maximum MSE (i.e., their worst-case performance) or based on averaged MSE (i.e., Bayes risk).

Summary

Estimators you'll encounter in Stat 410 take the following forms:

1. $\hat{\theta}$ = the sample mean $\bar{X}$ or linear functions of the sample mean, e.g., $2\bar{X}$.
   - $\mathbb{E}(a\bar{X}) = a \cdot \mu$ (usually unbiased).
   - $\text{Var}(a\bar{X}) = a^2 \frac{\sigma^2}{n}$.

2. $\hat{\theta}$ = the average of a transformation of X_i's, e.g., $\frac{1}{n}\sum_{i=1} \log X_i$ or $\frac{1}{n}\sum_{i=1} X_i^2$.
   - Define $Y_i = \log X_i$ then $\hat{\theta} = \frac{1}{n}\sum_{i=1} \log X_i = \bar{Y}$.
   - Find the distribution of Y_i and then you are back to the previous case.

3. $\hat{\theta} = g(\bar{X})$, a function of the sample mean, e.g., $1/\bar{X}$.
   - Check the sign of the second derivative of g in the range of possible values of X.
   - Use Jensen's inequality to show $\hat{\theta}$ is biased.
   - Usually you won't be asked to compute the bias and variance.

4. $\hat{\theta}$ = order statistics, e.g., $Y_1 = \min_i X_i$ or $Y_n = \max_i X_i$.
   - Find the distribution of Y_1 or Y_n .
   - Compute the mean (usually biased) and variance. For variance, use formula $\mathbb{E}Y_n^2 - (\mathbb{E}Y_n)^2$.

Appendix

(Feel free to ignore the materials in the Appendix.)

- Let $x_1, \ldots, x_n$ be a sequence of numbers, find the value a that minimizes

$$g(a) = \sum_{i=1}^{n} |x_i - a|.$$

It is equivalent to write the objective function as

$$g(a) = \sum_{i=1}^{n} |x_{(i)} - a|,$$

where $x_{(1)} < \cdots < x_{(n)}$ are the order statistics of x_i's. It's okay to assume that all the x_i's are different (i.e., no ties), since they are usually random samples from a continuous distribution.

Recall this result: suppose $g(x) = |x|$, then $g'(x) = 1$ if $x > 0$, -1 if $x < 0$, and $g'(0)$ does not exist. So the derivation of g is not well defined at x_i's.

Suppose $a \leq x_{(1)}$, then

$$g(a) = \sum_{i=1}^{n} (x_{(i)} - a) = \sum_{i} x_i - na,$$

which is a decreasing function, so its minimal is achieved at $a = x_{(1)}$. Similarly, we can check the case when $a \geq x_{(n)}$. We can conclude that it suffices to find the optimal value of a in the data range, $[x_{(1)}, x_{(n)}]$.

Assume we have an odd number of samples, i.e., $n = 2m + 1$.

- When $a \in [x_{(1)}, x_{(m+1)}]$, $g(a)$ is a decreasing function;
- When $a \in [x_{m+1}, x_n]$, $g(a)$ is an increasing function.

Therefore the minimal of $g(a)$ is achieved when $a = x_{(m+1)}$, the median of $(x_1, \ldots, x_n)$.

Assume we have an even number of samples, i.e., $n = 2m$.

- When $\theta \in [x_{(1)}, x_{(m)}]$, $g(a)$ is a decreasing function;
- When $a \in [x_{(m+1)}, x_{(n)}]$, $g(a)$ is an increasing function;
- When $a \in [x_{(m)}, x_{(m+1)}]$, $g(a)$ is a constant.

So the minimizer of $g(a)$ is any value in the interval $[x_{(m)}, x_{(m+1)}]$, the sample median of the n data points, which is not unique.

- $\max_{p_1,\ldots,p_m} \sum_{j=1}^m n_j \log p_j$ where $0 \le p_j \le 1$ and $\sum_j p_j = 1$ and $n_j \ge 0$, is achieved by setting $p_j = n_j/n$ where $n = n_1 + \cdots + n_m$.

  This is a constrained optimization problem, which, of course, can be solved by using tools like the Lagrange multiplier. Next I'll give a simple derivation based on the non-negativity of the Kullback-Leibler divergence.

  Scale the objective function by $(-\frac{1}{n})$ and look for the minimal. We have

$$\sum_{j=1}^m \frac{n_j}{n} \log \frac{1}{p_j} = \sum_{j=1}^m \frac{n_j}{n} \log \frac{n_j/n}{p_j} - \sum_{j=1}^m \frac{n_j}{n} \log \frac{n_j}{n},$$

  where the second term (on the right) has nothing to do with $(p_1, \ldots, p_m)$, and the first term is the Kullback-Leibler divergence between two multinomial distributions whose minimal is achieved by setting

$$\hat{p}_1 = \frac{n_1}{n}, \cdots, \quad \hat{p}_m = \frac{n_m}{n}.$$